Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis

Year: 2023 | Citations: 174 | Authors: Hubert Siuzdak

Abstract

Recent advancements in neural vocoding are predominantly driven by Generative Adversarial Networks (GANs) operating in the time-domain. While effective, this approach neglects the inductive bias offered by time-frequency representations, resulting in reduntant and computionally-intensive upsampling operations. Fourier-based time-frequency representation is an appealing alternative, aligning more accurately with human auditory perception, and benefitting from well-established fast algorithms for its computation. Nevertheless, direct reconstruction of complex-valued spectrograms has been historically problematic, primarily due to phase recovery issues. This study seeks to close this gap by presenting Vocos, a new model that directly generates Fourier spectral coefficients. Vocos not only matches the state-of-the-art in audio quality, as demonstrated in our evaluations, but it also substantially improves computational efficiency, achieving an order of magnitude increase in speed compared to prevailing time-domain neural vocoding approaches. The source code and model weights have been open-sourced at https://github.com/gemelo-ai/vocos.